Finding Higher Order Motifs under the Levenshtein Measure

Ezekiel F. Adebiyi and Tinuke Dipe
Department of Mathematics and Computer Science
University of Ilorin, Ilorin
Nigeria
adebiyi@informatik.uni-tuebingen.de, dipetinuke@yahoo.com

Abstract

We study the problem of finding higher order motifs under the levenshtein measure, otherwise known as the edit distance. In the problem set-up, we are given N sequences, each of average length n, over a finite alphabet Σ and thresholds D and q, we are to find composite motifs that contain motifs of length P (these motifs occur with atmost D differences) in $1 \leq q \leq N$ distinct sequences.

Two interesting but involved algorithms for finding higher order motifs under the edit distance was presented by Marsan and Sagot[7]. Their second algorithm is much more complicated and its complexity is asymptotically not better. Their first algorithm runs in $O(M \cdot N^2 n^{1+\alpha \cdot p \cdot pow(\epsilon)})$, where $p \geq 2$, $\alpha > 0$, $pow(\epsilon)$ is a concave function that is less than 1, $\epsilon = D/P$ and M is the expected number of all monad motifs. We present an alternative algorithmic approach also for Edit distance based on the concept described in [3, 4]. The resulting algorithm is simpler and runs in $O(N^2 n^{1+p \cdot pow(\epsilon)})$ expected time.

1. Introduction

Pattern discovery in unaligned DNA sequences is a fundamental problem in computational biology with important applications in finding regulatory signals. Existing approaches on finding patterns focus on monad patterns, otherwise known as common motifs[9] or simply motifs[8], that correspond to relatively short contiguous strings[1]. A new problem of finding composite motifs, otherwise known as structured motifs[12] or higher order motifs[11], arises, when the number of monad motifs, p, that participates in a biological process is greater than one. The relative position of each monad motif is now important and is not random but sterically defined. In eukaryotic, these monad motifs may interact with one another. Recently, increasing efforts have been made to tackle this problem, but with respect to hamming distance[11, 6]. It is known[1, 8, 10] that sequences of organisms exist, where insertion and deletions are important to find regulatory signals. It is this challenge that we consider in this project. We consider here the simple unit measure, whereby, the Edit distance between two strings $S_1[1..i]$ and $S_2[1..j]$, $\delta(i,j)$ is defined as follows:

$$\delta(i,j) = \min[\delta(i-1,j)+1, \delta(i,j-1)+1, \delta(i-1,j-1)+t(i,j)],$$

where $t(i,j)$ is 0, if $S_1(i) = S_2(j)$ else 1. The concept of a model is used in the two algorithms described below. A model of length P is a pattern over Σ^P. A valid model, also known as a consensus, is a model that present a single representation for all reoccurrence of a motif.

2. Two Suffix tree-based Algorithms

The algorithms we will implement in this project are built using the Generalised Suffix tree (GST), an hybrid of a suffix tree (ST)[1]. Marsan and Sagot[7] considered theoretically the finding of higher order motifs under the edit distance and presented two algorithms. Their second algorithm is much more complicated and its complexity is asymptotically not better.

Theorem 1 The running time of their first algorithm under the edit distance is $O(M \cdot N^2 n^{1+\alpha \cdot p \cdot pow(\epsilon)})$, where $p \geq 2$, $\alpha > 0$, $pow(\epsilon)$ is a concave function that is less than 1, $\epsilon = D/P$ and M is the expected number of all monad motifs.

M is omitted in the complexity analysis given in [7], but M is not constant. Under the hamming distance, Buhler and Tompa[5] estimated $M = |\Sigma|^P (1-(1-P_D)^{n-P+1})^q$, where P_D is the probability that a given motif of length P, occurs with upto D substitutions at a given position of a random sequence. We showed in [1] that under the edit distance, $P_D < 2N^{\alpha \cdot (pow(\epsilon)-1)}$.

The naive method proposed by Marsan and Sagot[12, 7] consisted basically of extracting and storing all valid motifs

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

of length P and verifying which pairs of such motifs could represent valid composite motifs, given an interval of distance. Observe that the motifs that formed a composite motif, must not come from the same model but must be on the same q sequences with defined distance between each other. Therefore, the extracting of motifs and verifying of a posteriori of which pairs of such motifs could represent valid composite motifs, can be performed simultaneously. This is the key observation in the construction of our alternative algorithmic approach[1]. It begins by finding motif under the edit distance (using our motifs extraction algorithm in [4]), and then simultaneously verify if it form composite motifs with another existing motifs.

Theorem 2 The running time of our algorithm under the edit distance is $O(N^2 n^{1+p \cdot pow(\epsilon)})$, where $p \geq 2$, $\epsilon = D/P$ and $pow(\epsilon)$ is a concave function less than 1.

3. Conclusion

We have presented and considered the theoretical efficiency of two algorithms for finding higher order motifs under the edit distance. More indepth work will be done in [2] as was carried out in [4]. There, we will compare their practical efficiency and effectiveness using real and simulated data.

References

[1] E. Adebiyi. *Pattern Discovery in Biology and Strings Sorting: Theory and Experimentation*. Shaker Verlag, Aachen, 2002.

[2] E. Adebiyi and T. Dipe. Finding higher order motif under the levenshtein measure. *In preparation*, 2003.

[3] E. Adebiyi, T. Jiang, and M. Kaufmann. An efficient algorithm for finding short approximate non-tandem repeats (extended abstract). *ISMB 2001 Bioinformatics*, 17((Suppl. 1)):S5–S13, July 2001.

[4] E. Adebiyi and M. Kaufmann. Extracting common motifs under the levenshtein measure: Theory and experimentation. *WABI*, 2002.

[5] J. Buhler and M. Tompa. Finding motifs using random projections. *RECOMB*, 2001.

[6] E. Eskin. Finding composite regulatory patterns in dna sequences. *manuscript*, 2001.

[7] L. Marsan and M. F. Sagot. Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification. *RECOMB 2000*.

[8] P. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in dna sequences. *ISMB*, pages 269–278, 2000.

[9] M.-F. Sagot. Spelling approximate repeated or common motifs using a suffix tree. *LNCS*, 1380:111–127, 1998.

[10] O. Sand. Topological representation of transcriptional regulatory regions. *manuscript*, 2002.

[11] S. Sinha. Composite motifs in promoter regions of genes: models and algorithms. *General Report*, 2002.

[12] A. Vanet, L. Marsan, A. Labigne, and M.-F. Sagot. Inferring regulatory elements from a whole genome. an analysis of helicobacter pylori σ^{80} family of promoter signals. *J. Mol. Biol.*, 297:335–353, 2000.